Discovering Common Substructures in Proteins

Ye, Jieping[1], Janardan, Ravi[*1], Liu, Songtao[2]
[1]University of Minnesota, Minneapolis, MN, USA; [2]University of Alberta, Edmonton, Alberta, Canada

Three-dimensional structure plays a central role in research directed towards understanding evolutionary and functional relationships among proteins. For instance, it is well-known that structural information is better conserved than sequence information in the evolution of proteins, hence can be used in the construction of phylogenetic trees. Protein-protein interactions are governed in large part by the shape, location, and composition of the active sites. The assignment of proteins to fold families is accomplished via structural analysis. The need for effective structural analysis techniques has increased with the rapid growth in the number of three-dimensional structures available now in repositories such as the Protein Data Bank.

The focus of this project is on protein structure alignment. Informally, given a collection of proteins, represented by their C_α-backbone structures, the goal is to find a superposition of these structures in space, via rigid motions, such that large matching structures are revealed. First, an efficient algorithm is developed for aligning two proteins (*pairwise structure alignment*) such that the aligned structures have large size and aligned atoms are within a user-specified distance of each other. The algorithm uses a representation of the backbones that is independent of their relative orientations in space and applies dynamic programming to this representation to compute an initial alignment, which is then refined iteratively. This algorithm is then incorporated within an algorithm for aligning multiple protein structures (*multiple structure alignment*) such that the sum of the pairwise distances between the aligned structures is small. The algorithm also generates from the given collection, a consensus (pseudo)protein that represents the entire set. The algorithm picks an input protein as an initial consensus and uses an adaptation of the well-known *center-star* method for multiple sequence alignment to compute a correspondence between the different proteins. It then derives from this correspondence a set of transformations to align the structures and generate simultaneously a new consensus. This process is iterated until the sum of pairwise distances converges. This approach is based on an interesting result that allows the sum of all pairwise distances to be represented compactly as distances to the consensus. Experimental results on several protein families show that the two algorithms are competitive with known ones.

This effort is sponsored, in part, by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.